

Dear sir,

7 datasets are provided for this task. I find out the most relevant datasets after understanding the data model. Before creating a final dataset, I did data cleaning of all datasets.

During data cleaning, I found out that only the “Reaction” dataset has null values. There are 980 null values present in the “Type” column of the dataset. Null values are removed because it is an important column for further analysis. In the “Session” dataset, some device names are not correct. These names are converted into “Other”. The “category” column in the “content” dataset has some values in the quote which are the same as those that are not in the quote. So, those values were replaced. The user ID column in the “Reaction” dataset was removed.

Now, the dataset is ready to join. I used an SQL server for joining the different tables.